International Journal of Innovative
Computing, Information and Control
Volume **7**, Number **8**, August **2011**

 ISSN 1349-4198
pp. **5027–5043**

DETECTION OF LINE-SYMMETRY CLUSTERS

Yi-Zeng Hsieh[1], Mu-Chun Su[1], Chien-Hsing Chou[2] and Pa-Chun Wang[3]

[1]Department of Computer Science and Information Engineering
National Central University
No. 300, Jhongda Rd., Jhongli City, Taoyuan County 32001, Taiwan
{ eizon; muchun }@csie.ncu.edu.tw

[2]Department of Electrical Engineering
Tamkang University
No. 151, Yingzhuan Rd., Danshui Dist., New Taipei City 25137, Taiwan
chchou@mail.tku.edu.tw

[3]Quality Management Center
Cathay General Hospital
No. 280, Renai Rd., Sec. 4, Taipei, Taiwan
drtony@tpts4.seed.net.tw

Received May 2010; revised September 2010

Abstract. *Many real-world and man-made objects are symmetry. Therefore, it is reasonable to assume that some kinds of symmetry may exist in data clusters. The most common type of symmetry is line symmetry. In this paper, we propose a line symmetry distance measure. Based on the proposed line symmetry distance, a modified version of the K-means algorithm can be used to partition data into clusters with different geometrical shapes. Several data sets are used to test the performance of the proposed modified version of the K-means algorithm incorporated with the line symmetry distance.*
Keywords: Cluster analysis, Clustering algorithm, Symmetry, Distance measure

1. **Introduction.** Cluster analysis is a tool to explore the underlying structure of a given data set. It has been applied in many applications [1-4]. While it is easy to consider the idea of a data cluster on a rather informal basis, it is very difficult to give a formal and universal definition of a cluster. In the field of cluster analysis, there are two major difficulties encountered in clustering data. First of all, we usually do not know the number of clusters in a high-dimensional data set in advance. Basically, there are three different approaches to the determination of the cluster number of a data set. The first approach is to use a certain global validity measure (e.g., the Dunn's separation measure [5,6], the Bezdek's partition coefficient [7], the Xie-Beni's separation measure [8], Davies-Bouldin's measure [9], the Gath-Geva's measure [10], the CS measure [11]) to validate clustering results for a range of cluster numbers. The second approach is based on the idea of performing progressive clustering [12-18]. The third approach is the projection-based approach. Projection algorithms allow us to visualize high-dimensional data as a two-dimensional or three-dimensional scatter plot [15-21]. Even if we are able to solve the problem of estimating the number of clusters, the next crucial problem which needs to be solved is the determination of the distance measure. The geometric shapes of clusters vary from a data set to a data set. Even in the same data set, each cluster may have its own geometrical property. Different distance measures may lead to different geometrical-shaped clusters. For example, while the popular Euclidean distance leads to the detection